

Towards Robust Mathematical Reasoning

Thang Luong[◇], Dawsen Hwang^{*}, Hoang H. Nguyen^{*,†},
Golnaz Ghiasi^{*}, Yuri Chervonyi^{*}, Insuk Seo^{*,†}, Junsu Kim^{*},

Garrett Bingham, Jonathan Lee, Swaroop Mishra[†], Alex Zhai, Clara Huiyi Hu, Henryk Michalewski,
Jimin Kim[†], Jeonghyun Ahn[†], Junhwi Bae[†], Xingyou Song, Trieu H. Trinh, Quoc V. Le, Junehyuk Jung[◇]

Google DeepMind, [◇] Corresponding authors, ^{*}Core and equal contributors

[†]Work previously conducted under Google DeepMind

Abstract

Finding the right north-star metrics is highly critical for advancing mathematical reasoning capabilities of foundation models, especially given that existing evaluations are either too easy or only focusing on getting correct short answers. To address these issues, we present *IMO-Bench*, a suite of advanced reasoning benchmarks that specifically targets the level of the International Mathematical Olympiad (IMO), the most prestigious venue for young mathematicians. *IMO-AnswerBench* first tests models on 400 diverse Olympiad problems with verifiable short answers. *IMO-ProofBench* is the next-level evaluation for proof-writing capabilities, which includes both basic and advanced IMO problems as well as detailed grading guidelines to facilitate automatic grading. These benchmarks played a crucial role in our historic achievement of the gold-level performance at IMO 2025 with Gemini Deep Think (Luong and Lockhart, 2025). Our model achieved 80.0% on *IMO-AnswerBench* and 65.7% on the advanced *IMO-ProofBench*, surpassing the best non-Gemini models by large margins of 6.9% and 42.4% respectively. We also showed that autograders built with Gemini reasoning correlate well with human evaluations and construct *IMO-GradingBench*, with 1000 human gradings on proofs, to enable further progress in automatic evaluation of long-form answers. We hope that *IMO-Bench* will help the community towards advancing robust mathematical reasoning and release it at <https://github.com/google-deepmind/superhuman/imobench>.

1 Introduction

The field of artificial intelligence, particularly large language or foundation models, has demonstrated

[◇]Correspondence: thangluong@google.com, junehyuk@google.com. [†]External affiliations: Georgia Institute of Technology (Hoang Nguyen), Seoul National University (Insuk Seo, Junsu Kim, Jimin Kim), Microsoft (Swaroop Mishra), Massachusetts Institute of Technology (Jeonghyun Ahn, Junhwi Bae), Brown University (Junehyuk Jung).

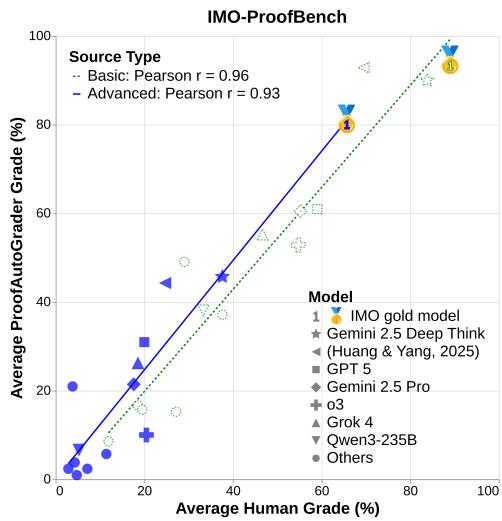


Figure 1: *IMO-ProofBench*, a benchmark in *IMO-Bench*, for measuring proof-writing capabilities. We demonstrated high correlations between human and automatic evaluations on a variety of public models, including our IMO-gold model. See §3 and §5.3 for more details.

remarkable progress in mathematical reasoning capabilities. Many popular benchmarks such as GSM8K (Cobbe et al., 2021), Hendrycks’ MATH (Hendrycks et al., 2021), and the recently popular AIME (for American Invitational Mathematics Examination) have approached saturation, limiting their usefulness in differentiating model performances. The problems in these datasets often rely on a limited set of techniques and do not always require the deep, multi-step reasoning needed to truly evaluate AI mathematical reasoning. Indeed, relying on final answer matching, even in recent benchmarks such as FrontierMath (Glazer et al., 2024) and Humanity’s Last Exam (Phan et al., 2025), is not entirely reliable. It could lead to AI systems that are good at guessing answers but do not exhibit robust reasoning skills.

To address these shortcomings, we propose *IMO-Bench*, a suite of benchmarks that focus on robust

Benchmark	Size	Task
<i>IMO-AnswerBench</i>	400	Get the right answer
<i>IMO-ProofBench</i>	60	Write a rigorous proof
<i>IMO-GradingBench</i>	1000	Grade a proof

Table 1: Benchmarks in the *IMO-Bench* suite.

reasoning at the level of the International Mathematical Olympiad (IMO), the world’s most celebrated arena for young mathematicians. The IMO is selected due to its notoriously difficult problems, which require not only rigorous multi-step reasoning but also a high degree of novelty, going beyond the simple application of known formulas. Such characteristics make IMO an excellent testbed for assessing reasoning capability. *IMO-Bench* covers three different tasks as summarized in Table 1 and all problems were vetted by a panel of IMO medalists¹ and mathematicians.

The first benchmark, *IMO-AnswerBench*, consists of 400 problems with verifiable answers carefully chosen from from past Olympiad competitions and then altered by experts to avoid memorization. Problems were chosen from a variety of topics whose solutions require different problem solving techniques to ensure a diverse representation of topics, ideas, and domain knowledge as illustrated in Figure 2.

The second benchmark, *IMO-ProofBench*, consists of 60 problems of varying difficulty levels, similar to those found at the IMO. While some problems have short answers, all require models to generate complete proofs. The benchmark is divided into two subsets, *basic* and *advanced*, each with 30 problems. While the basic set covers difficulty levels from pre-IMO up to IMO-Medium, problems in the advanced set are up to IMO-hard level and comprises of 5 complete IMO sets, 3 of which are novel. We designed this benchmark to shift the community’s focus from final answers to the proofs themselves, enabling a more rigorous assessment of AI reasoning processes. To ensure consistent evaluation, we include detailed grading schemes suitable for both human experts and automated systems. Figure 1 provides an early look into the potential of automatic graders for proofs.

These two benchmarks played a crucial role in the development of our Gemini Deep Think, leading to the historic achievement of the gold-level performance at IMO 2025 (Luong and Lockhart,

¹Together, they won 10 gold and 5 silver IMO medals.

2025). Our IMO-gold model achieved an accuracy of 80.0% on IMO-AnswerBench by automatic evaluation, surpassing the best non-Gemini model and the best open-weight model by a large margin of 6.9% and 19.2% respectively. The advanced IMO-ProofBench is much more challenging. Our IMO-gold scored 65.7%, whereas the best non-Gemini and the best open-weight models performed poorly with only 23.3% and 7.1% accuracy according to human evaluations. Furthermore, we demonstrate that automated graders for both answers and proofs, built upon Gemini 2.5 Pro, achieve high correlation with expert human evaluations.

Last but not least, we introduce *IMO-GradingBench*, a benchmark of 1000 solutions to problems in the advanced IMO-ProofBench, together with grades from human experts. This resource is designed to foster progress in the automatic evaluation of long-form answers. We release² *IMO-Bench* to the community and hope that it will spur further research towards advancing robust mathematical reasoning.

2 IMO-AnswerBench

2.1 Problem Selection

400 math problems were handpicked from various national, regional and international Olympiad contests, spanning across four categories (Algebra, Combinatorics, Geometry, and Number Theory). For each category, the benchmark contains 100 problems across four levels of difficulty: pre-IMO (middle school or pre-Math Olympiad problems), IMO-Easy (equivalent to Problem 1 or Problem 4 at the IMO), IMO-Medium (equivalent to Problem 2 or Problem 5 at the IMO) and IMO-Hard (equivalent to Problem 3 or Problem 6 at the IMO or post-Math Olympiad problems). The difficulty breakdown for each category is listed in Table 2.

Category	Pre-IMO	IMO-Easy	IMO-Medium	IMO-Hard
Algebra	11	46	32	11
Combinatorics	4	19	31	46
Geometry	13	44	32	11
Number Theory	2	20	31	47

Table 2: Difficulty breakdown for *IMO-AnswerBench*.

Problems with short answers were chosen so the correctness of a model’s output can be quickly and reliably determined. Given the proof-heavy nature of many math Olympiad problems, we perform an

²<https://github.com/google-deepmind/superhuman/imobench>

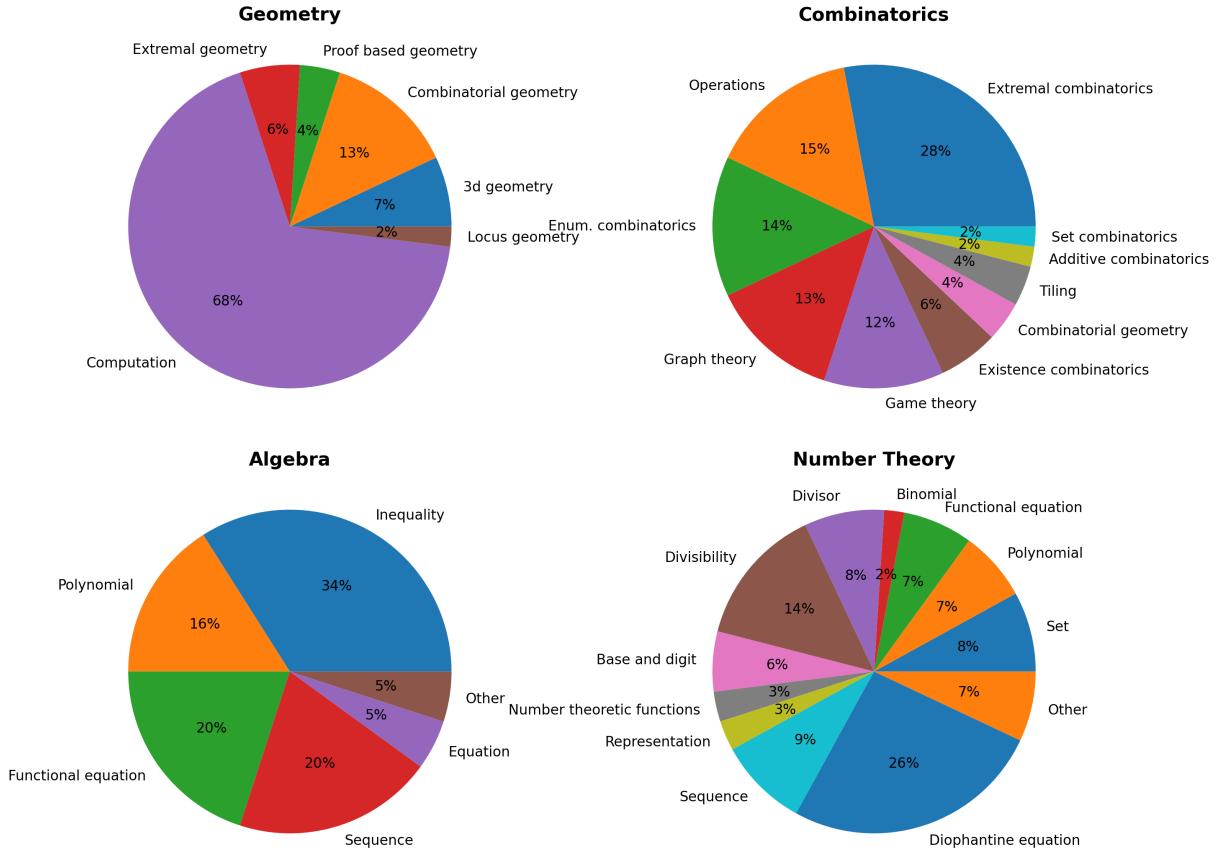


Figure 2: Topic distribution by category in *IMO-AnswerBench*. Number Theory and Combinatorics have the most topics which reflect the broad knowledge required to solve these problems while Geometry is mostly skewed towards angle and sidelength computation problems due to the nature of the short answer benchmark.

additional reformulation step for certain examples. This adjustment ensures that each problem yields a clear and nontrivial short answer, thereby reducing ambiguity during solving and verification and confirming that models utilize nontrivial reasoning. We detail specific examples in A.4.

2.2 Problem Robustification

To avoid data memorization, an additional step of problem modification is done via paraphrasing, changing the name of objects in the problem (such as changing point names for geometry problems), reformulating, modifying numerical values and/or adding distractors to the problem. This process is done either manually or automatically using language models. We highlight some examples in Table 8 and detail below.

One example is an algebra problem from Austria Math Olympiad 2017. The problem is modified by making the substitution $x = a+b-c$, $y = b+c-a$, and $z = c+a-b$ for positive real numbers x, y, z with a, b , and c being the lengths of the sides of some triangle to obtain the modified problem in

the *Robustified* column. This modification uses the knowledge that a, b , and c are lengths of a triangle if and only if they satisfy the triangle inequalities $a + b > c$, $a + c > b$, and $b + c > a$.

Another example is a combinatorics problem from USA TST 2005. From the original statement, the problem is modified using several techniques such as modifying numerical values (by assigning a specific value to the variable n so that it is harder to guess the pattern), adding distractors (by introducing a function or variables that are not relevant to the problem), and adding a layer of challenge that could confuse the models.

Experts also reformulated original problems into equivalent ones with completely different expressions. One such example is the Czech-Slovak Math Olympiad 2017 problem, where we obtain a robustified problem by transforming the governing equation and changing the objective from finding all possible values of k to finding all even integers d such that the number of solutions is even.

2.3 Answer autograder

Even for the problems with short answers, automatic answer verification presents a few substantial challenges. The difficulty arises from two main issues: (1) ensuring that model outputs adhere to a parsable format and (2) evaluating semantically equivalent but syntactically different expressions.³ To circumvent this issue, benchmarks such as FrontierMath (Glazer et al., 2024) select problems with only numerical answers or mathematical objects that can be expressed as SymPy objects. However, this approach narrows the scope of evaluable problems and reduces robustness of the benchmark to minor formatting or syntax errors.

To address these limitations, we incorporate an approach of prompting a large language model as an automated verifier for the model answers on *IMO-AnswerBench*. We name this language model component as *AnswerAutoGrader*. The *AnswerAutoGrader* is built by prompting the public Gemini 2.5 Pro model to extract the final answer from the generated solution and assess its correctness against the ground truth (See A.5 for the full prompt). This method allows much more flexibility in acceptable answer formats and improves the overall robustness of our benchmark. As we demonstrate in Section 5.1, *AnswerAutoGrader*'s performance is nearly identical to that of human evaluators, validating its use for future public usage and also for reporting the results in this work.

3 Going Beyond Short Answers with IMO-ProofBench

While the final answer accuracy provided by *IMO-AnswerBench* offers a valuable metric for measuring mathematical abilities, it is insufficient for a comprehensive assessment of mathematical reasoning. A final answer can be correct while the full solution contains flawed reasoning. Furthermore, many IMO-level competition problems do not come with a final short answer. Even in cases where a short answer exists, guessing the correct short answer is often significantly easier than rigorously deriving the solution.

IMO-ProofBench is designed to evaluate the ability of AI models to construct comprehensive and valid mathematical arguments. This benchmark consists of 60 proof-based problems, curated to

³For example, given the ground truth answer " $(-\infty, -4) \cup (-4, \infty)$ ", the answer "all real numbers except -4" should also be graded as correct.

mirror the kinds of problems found in the IMO. While some problems may have concise numerical answers, models are only given credit if they produce correct and relevant reasoning steps. This benchmark is essential for assessing an AI's underlying reasoning process, its ability to apply mathematical principles, and its capacity to formulate coherent and logical arguments.

3.1 Benchmark setup

The benchmark is divided into two subsets: a *basic* set covering pre-IMO to IMO-Medium difficulty levels, and an *advanced* set featuring novel, highly challenging problems simulating complete IMO examinations, up to IMO-Hard level.

The basic problem set primarily consists of rephrased versions of existing problems. Since standard IMO problems may be too challenging for most of current models, the basic set is designed to assess models in their early stages of development. Sufficiently strong performance on the basic set would justify progression to the advanced set.

The advanced problem set features 30 problems in the style and difficulty of the IMO. The collection includes 18 novel problems crafted by IMO medalists, alongside 12 problems from recent top-tier competitions: 6 robustified from IMO 2024 and 6 directly from USAMO 2025. Table 10 provides examples of such robustified problems.

IMO-ProofBench uses an evaluation framework designed for both simplicity and precision. We provide a primary grading guideline with four ratings (*Correct*, *Almost*, *Partial*, *Incorrect*) as detailed in Table 3. While this rubric offers a clear and consistent baseline, we do not restrict our expert evaluators to these four values. To allow for more nuanced assessments, human experts are empowered to use their own judgments to assign any integer score from 0 to 7 for each problem.

Category	IMO Points	Solution quality
Correct	7	Fully correct, rigorous, and complete
Almost	6	Almost correct, minor errors
Partial	1	Mostly incorrect, some relevant results
Incorrect	0	Completely incorrect or irrelevant

Table 3: Our simplified IMO ratings.

3.2 Proof Autograder

While human expert evaluation remains the gold standard for mathematical proofs, its cost and time intensity limit scalable research. To address this,

we built *ProofAutoGrader*, an automatic grader for *IMO-ProofBench*. The autograder leverages Gemini 2.5 Pro, providing it with a prompt containing the problem statement, the candidate solution, a reference solution, and specific grading guidelines (see Appendix B.5).

Automatic evaluation for informal proofs is a highly intricate task, and current systems are not yet a perfect substitute for human experts—a key distinction from *AnswerAutoGrader*, whose purpose is primarily format matching. For this reason, all primary results in this paper are based on expert human evaluation to ensure all results are absolutely correct. Nevertheless, as we demonstrate in Section 5.3, we prove our autograder can be a reasonable proxy, establishing it as a reasonable tool for the community to assess future models on *IMO-ProofBench*.

4 IMO-GradingBench

While *IMO-ProofBench* evaluates proof-writing abilities, it is equally important to assess models in terms of their ability to evaluate the correctness of given solutions. This capability is crucial for developing reliable automated grading systems and improving general mathematical reasoning. This is the focus of *IMO-GradingBench*.

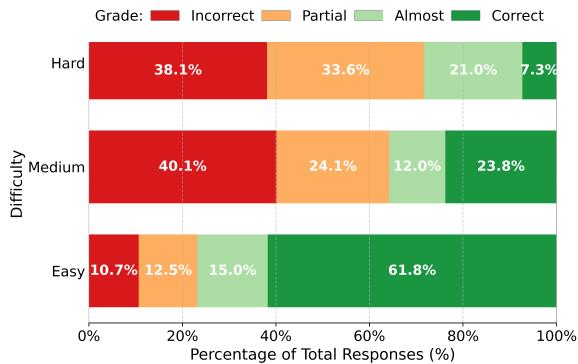


Figure 3: Grade distribution for solutions in *IMO-GradingBench* by difficulty levels (IMO-Hard, IMO-Medium, IMO-Easy).

As part of our IMO effort (Luong and Lockhart, 2025), we have benchmarked extensively many internal models on the advanced set of *IMO-ProofBench* using human evaluations, which has been tremendously instrumental for our rapid progress leading up to IMO 2025. These human gradings later have led to the creation of *IMO-GradingBench* with 1000 examples, each containing a problem statement, a proposed solution, and

its human-assigned grade (on a 0–7 scale). To reduce noise from fine-grained scoring, we frame the evaluation as a four-way classification by mapping the given IMO points to the labels (*Correct*, *Almost*, *Partial*, *Incorrect*) as detailed in Table 3.

To ensure a robust evaluation, the dataset has been balanced with a roughly equal number of examples per category. Figure 3 illustrates that when problems are grouped by their IMO difficulties, a clear trend emerges. The proportion of correct and almost solutions decreases as the intended difficulty moves from IMO-easy to IMO-hard, while the proportion of incorrect and partial solutions increases. This confirms that the grading distribution of *IMO-GradingBench* aligns with its assigned difficulty levels. See further discussions in Section C.1.

5 Results

We evaluate *IMO-Bench* on a wide variety of publicly available models: Claude Opus 4 (20250514), Claude Sonnet 4 (Anthropic, 2025), DeepSeek V3 (DeepSeek, 2025b), DeepSeek R1 (DeepSeek, 2025a), Kimi-K2-Instruct (Moonshot AI, 2025), Qwen3-235B (A22B-Instruct-2507- tput) (Qwen Team, 2025), o3 (2025-04-16), o4-mini (*high reasoning*) (OpenAI, 2025b), GPT-5 (2025-08-07) (OpenAI, 2025a), Gemini 2.5 Pro (Google DeepMind, 2025), Gemini 2.5 Deep Think (Deep Think team, 2025), Gemini Deep Think (*IMO Gold*) (Luong and Lockhart, 2025), Gemini 2.5 Pro with (Huang & Yang, 2025) (Huang and Yang, 2025), Grok 4 (0709) (xAI, 2025).

Note that since Gemini 2.5 Pro with (Huang & Yang, 2025) is an agentic framework rather than a single model call, Appendix B.3 contains further details on its implementation.

5.1 IMO-AnswerBench with AnswerAutoGrader

Results for *IMO-AnswerBench* are in Table 4. Accuracy was determined by *AnswerAutoGrader*, which extracts the final answer from the model response and assesses its semantic equivalence to the ground truth. Our Gemini Deep Think (*IMO Gold*) model achieved an overall accuracy of 80.0%, surpassing the best non-Gemini model (Grok 4) by 6.9% and the best open-weight model (DeepSeek R1) by 19.2%. Latest models such as Kimi-K2-Instruct and GPT-5 are still struggling with overall accuracy of only 45.8% and 65.6% respectively.

Across the four categories of Algebra, Combi-

Model	Query date	Algebra	Combinatorics	Geometry	Number Theory	Overall
Claude Opus 4	2025-08-04	19.4%	20.0%	23.3%	26.6%	22.3%
Claude Sonnet 4	2025-08-06	20.6%	17.8%	26.0%	27.6%	23.0%
DeepSeek V3	2025-09-17	39.0%	26.0%	35.0%	48.0%	37.0%
Kimi-K2-Instruct	2025-09-17	45.6%	31.1%	49.3%	56.9%	45.8%
Qwen3-235B	2025-08-20	57.6%	37.5%	57.6%	62.3%	53.8%
DeepSeek R1	2025-09-17	65.0%	40.0%	73.0%	65.0%	60.8%
o3	2025-08-04	62.8%	43.0%	70.6%	68.0%	61.1%
GPT-5	2025-09-17	69.9%	46.4%	74.8%	71.2%	65.6%
o4-mini (<i>high reasoning</i>)	2025-08-04	71.3%	46.6%	78.4%	75.3%	67.9%
Gemini 2.5 Pro	2025-08-04	73.4%	48.0%	74.3%	77.1%	68.2%
Gemini 2.5 Deep Think	2025-08-20	78.0%	49.0%	83.0%	77.0%	71.8%
Grok 4	2025-08-06	75.5%	55.9%	80.1%	80.9%	73.1%
Gemini Deep Think (<i>IMO Gold</i>)	2025-09-17	85.0%	69.0%	88.0%	78.0%	80.0%

Table 4: Model accuracy on *IMO-AnswerBench*. Results are averaged over 8 runs, except for Gemini 2.5 Deep Think and Gemini Deep Think (*IMO Gold*) (single run). An evaluation of Grok 4 (*heavy*) on 2025-08-13 using multiple paid accounts was aborted due to significant instability (only 117/400 responses were received despite multiple, hour-long attempts), and thus its results are not reported.

		<i>AnswerAutoGrader</i>	
		0	1
Human	0	274 (99.6%)	1 (0.4%)
Grade	1	8 (1.5%)	517 (98.5%)

Table 5: Confusion matrix for the *AnswerAutoGrader* predictions against human grades for answers to *IMO-AnswerBench*. The solutions were generated by Gemini 2.5 Pro and o3. Absolute counts are shown, with row-wise percentages in parentheses.

natorics, Geometry, and Number Theory, models generally perform the worst in Combinatorics, potentially highlighting difficulties with advanced abstract reasoning. We also analyze the performances of models on the original problems, before robustification, summarized in Appendix Table 9. As anticipated, we find robustification leads to a consistent drop in performance across all models.

Lastly, we validate the reliability of *AnswerAutoGrader* by comparing it with expert human labels. As reported in Table 5, the autograder shows nearly perfect performance, achieving overall accuracy of 98.9% on the positive (correct) class.

5.2 IMO-ProofBench with Expert Evaluations

Model outputs on *IMO-ProofBench* were graded by human experts according to the guidelines described in Section B.2. Table 6 presents the results of this evaluation. Performance on the basic *IMO-ProofBench* varies significantly; while most models score below 60%, Gemini Deep Think (*IMO Gold*) achieves a high score of 89.0%. The perfor-

mances of other frontier models such as Qwen3-235B (33.3%) and GPT-5 (59.0%) show that there is still considerable room for improvements.

The advanced *IMO-ProofBench* proves to be a more significant challenge that all non-Gemini models score below 25%. Our Gemini Deep Think (*IMO Gold*) model achieved a score of 65.7% according to human evaluations, surpassing the best non-Gemini model (Grok 4 (*heavy*)) by a large margin of 42.4%. This represents a substantial leap in capability, but its distance from a perfect score indicates that even the strongest models have room for growth in sophisticated mathematical reasoning.

A breakdown of the advanced *IMO-ProofBench* reveals a significant performance disparity across problem types, suggesting potential overfitting in certain models. This trend is most evident with Grok 4 (*heavy*), which scores 76.2% on USAMO 2025 but only 11.1% on novel problems. Other models, including o3 (52.4% vs. 15.1%) and Gemini 2.5 Pro with (Huang & Yang, 2025) (52.4% vs. 17.5%), exhibit a similar, pronounced gap.

In contrast, Gemini Deep Think (*IMO Gold*) scored 69.0% on the USAMO and 61.1% on the novel sets, indicating that our model is more general (Deep Think team, 2025) and does not overfit to a particular math dataset. The low performances of latest frontier models such as GPT-5 and Grok 4 (*heavy*) on the advanced *IMO-ProofBench* underscore the difficulty of advanced mathematical reasoning and highlight the importance of rigorously checking the full details of model outputs for a complete understanding of their abilities.

Model	Query date	IMO-ProofBench		Advanced IMO-ProofBench Breakdown		
		Basic	Advanced	Novel	IMO 2024 [†]	USAMO 2025
Number of Problems		30	30	18	6	6
Claude Opus 4	2025-08-04	11.9%	2.9%	0.0%	2.4%	11.9%
DeepSeek V3	2025-09-16	18.6%	4.3%	6.3%	2.4%	0.0%
Kimi-K2-Instruct	2025-08-21	19.5%	7.1%	4.0%	2.4%	21.4%
Claude Sonnet 4	2025-09-17	27.1% ^{§1}	4.8% ^{§1}	6.4% ^{§1}	2.4%	2.4%
DeepSeek R1	2025-09-16	29.0%	3.8%	6.4%	0.0%	0.0%
Qwen3-235B	2025-08-21	33.3%	5.2%	7.1%	0.0%	4.8%
o4-mini (<i>high reasoning</i>)	2025-08-04	37.6%	11.4%	8.7%	7.1%	23.8%
Grok 4	2025-08-20	46.7%	18.6%	17.5%	16.7%	23.8%
o3	2025-08-04	54.8%	20.5%	15.1%	4.8%	52.4%
Gemini 2.5 Pro	2025-08-04	55.2%	17.6%	15.9%	7.1%	33.3%
GPT-5	2025-09-18	59.0%	20.0%	15.9%	33.3%	19.0%
Grok 4 (<i>heavy</i>)	2025-07-12	NA [‡]	23.3% ^{§3}	11.1% ^{§3}	7.1%	76.2%
Gemini 2.5 Pro with (Huang & Yang, 2025)	2025-07-14	69.5%	24.8%	17.5%	19.1%	52.4%
Gemini 2.5 Deep Think	2025-08-20	83.8%	37.6%	31.7%	40.5%	52.4%
Gemini Deep Think (<i>IMO Gold</i>)	2025-08-02	89.0%	65.7%	61.1%	76.2%	69.0%

Table 6: Expert evaluation results on the Basic and Advanced subsets of *IMO-ProofBench*. Scores are presented as a percentage of the total possible points for the problems in each respective subset, with each problem graded from 0–7 (as described in Section B.2). The Advanced *IMO-ProofBench* is further broken down by problem source.

[†]*Robustified IMO 2024 problem set, see Section 3.* [‡]*An attempt to query Grok 4 (heavy) on 2025-08-13 was unsuccessful due to model instability (only 5 of 30 problems responded with 3 attempts).* ^{§k}*Scores indicate that there were k problems that were treated as incorrect (a score of 0) because of query failures (for at least 3 times).*

5.3 Autograder for IMO-ProofBench

To assess the feasibility of using automatic graders for proofs, we apply *ProofAutoGrader* to the 14 public models (Table 6), previously graded by human experts on *IMO-ProofBench*. Figure 1 shows that the average grades from our automatic grader highly correlate with human grades, yielding high Pearson correlation coefficients of 0.96 and 0.93 on both basic and advanced problems respectively.

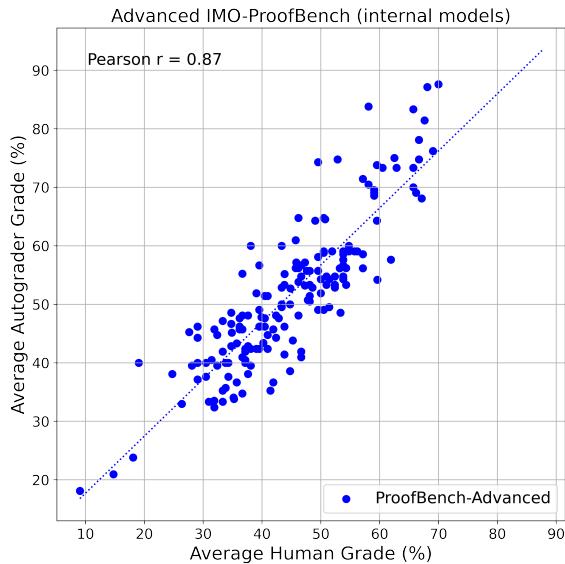


Figure 4: Correlation between *ProofAutoGrader* and human experts on the advanced *IMO-ProofBench*, evaluated over 170 internal models on our IMO-gold journey.

In addition, we also visualized, in Figure 4, the performance of *ProofAutoGrader* on 170 internal systems, developed as part of our IMO effort (Luong and Lockhart, 2025). On this larger pool, our automatic grader achieved a lower, but still reasonable Pearson correlation coefficient of 0.87.

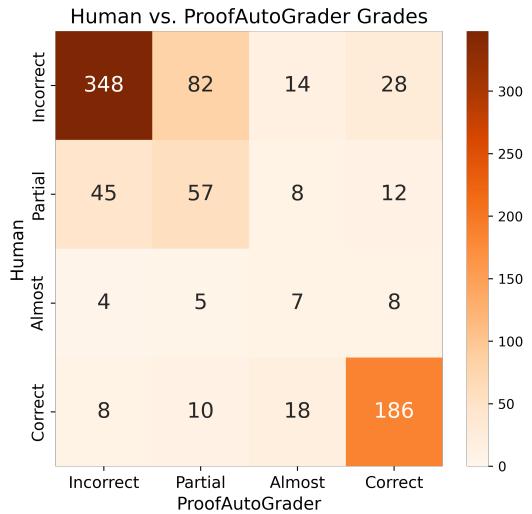


Figure 5: Per-solution confusion matrix for *ProofAutoGrader* vs. human expert grades, over 840 solutions generated by 14 public models (see Table 6).

To better understand the grading agreement, we visualize, in Figure 5, the confusion matrix of all human and automatic gradings on the 14 public models (for a total of 840 model solutions). We

observed that most common misclassifications happened between the *Incorrect* and *Partial* classes.

Detailed analysis in appendix B.6 further reveals that *ProofAutoGrader* occasionally still has weaknesses such as failures to identify high-level logical errors or being overly punitive for unconventional yet correct solutions. Therefore, while we hope that *ProofAutoGrader* can serve as a valuable tool for the community to evaluate models on *IMO-ProofBench*, we recommend couple its results with oversight and review from human.

5.4 IMO-GradingBench

The *IMO-GradingBench* measures the ability of models in assessing the quality of a proof when provided with only problem statements and model-generated solutions, without any reference solutions or specific grading guidelines. We measure model performances under two metrics:

1. *Accuracy* – human gradings on a 7-point scale are first converted to 4 categories (*Correct*, *Almost*, *Partial*, *Incorrect*) corresponding to 4 buckets (7, 6-4, 3-1, 0). The categorized human gradings are then compared with model-predicted categories.
2. *Mean Absolute Error (MAE)* – model-predicted categories are converted from (*Correct*, *Almost*, *Partial*, *Incorrect*) to IMO scores (7, 6, 1, 0) according to Table 3. We then compare with human grading ground truths on a 7-point scale.

Model	Accuracy \uparrow	MAE \downarrow
Gemini 2.5 Pro	44.3%	30.2%
o4-mini (<i>high reasoning</i>)	47.3%	25.2%
Gemini 2.5 Deep Think	52.5%	20.5%
o3	54.0%	20.2%
Gemini Deep Think (<i>IMO Gold</i>)	50.2%	18.4%

Table 7: *IMO-GradingBench* results: MAE calculated using simplified grades. Both model and human grades are mapped to categories (C, A, P, I) per Table 3, then to scores 7, 6, 1, 0. Avg. of 4 runs (Gemini 2.5 Deep Think and Gemini Deep Think (*IMO Gold*): 1 run).

Results for *IMO-GradingBench* are summarized in Table 7. In terms of accuracies, o3 achieved the highest performance of 54.0%. The low accuracies highlight the fact that this benchmark is quite challenging in predicting precise categories. The MAE (lower is better) accounts for the fact that different

categories are closer semantically, e.g., *Correct* vs. *Almost* and *Partial* vs. *Incorrect*. On this metric, Gemini Deep Think (*IMO Gold*) achieved the best MAE score of 18.4%, indicating that there is still significant room for improvement⁴.

Comparison with *ProofAutoGrader* – model performances on *IMO-GradingBench* are notably worse than what might be expected from the strong performance of *ProofAutoGrader*, in terms of Pearson correlation coefficients as reported in Section 5.3. This discrepancy stems from two key methodological distinctions.

1. First, *ProofAutoGrader* performance was measured on scores aggregated over 30 problems, which smooths out noise from individual grading variations, unlike the per-instance evaluation of *IMO-GradingBench*.
2. Second, the *IMO-GradingBench* evaluation provides models with minimal context—only the problem and the proposed solution; whereas for *ProofAutoGrader* on *IMO-ProofBench*, we additionally provide both reference solutions and grading guidelines.

These distinctions explain why *IMO-GradingBench* with per-instance, minimal-context evaluation is a challenging benchmark; whereas aggregated assessments by *ProofAutoGrader* on *IMO-ProofBench* can still yield robust model rankings.

6 Related Work

In recent years, harder reasoning math benchmarks have been proposed as performance on existing benchmarks becomes saturated. For example, Olympiad Bench (He et al., 2024) and Omni-MATH (Gao et al., 2024) contain questions at the Olympiad level across diverse domains, while Humanity’s Last Exam (HLE) (Phan et al., 2025) evaluates knowledge across many domains. Other benchmarks include Brainteaser (Han et al., 2025), which consists of long-form brainteaser puzzles, and Frontier Math (Glazer et al., 2024), which contains hard math questions and a hidden evaluation set. MiniF2F (Zheng et al., 2021) provides a benchmark for evaluating formal proofs around Olympiad-level difficulty. Reward Bench (Lambert et al., 2024) provides a benchmark to eval-

⁴Because of our simplified gradings (7, 6, 1, 0), the best possible grader will achieve a *golden* MAE of 3.9% on *IMO-GradingBench*, instead of 0%.

ate reward models. HARDMath (Fan et al., 2024) presents a challenging math benchmark containing applied mathematics problems that require analytical approximation techniques. The AlphaGeometry papers (Trinh et al., 2024; Chervonyi et al., 2025) provide benchmarks of 80 IMO and IMO Short-list Euclidean geometry problems from 2000 to 2024, written in a domain-specific language. In contrast, *IMO-Bench* provides a suite for evaluating advanced mathematical reasoning with short answer matching and rigorous proof evaluation in natural language across a wide variety of Math Olympiad areas.

As performance on math benchmarks continues to improve, robustness benchmarks have been introduced to evaluate potential overfitting and obtain better estimates of models’ true reasoning capabilities. These benchmarks have shown that simply perturbing benchmark questions is enough to significantly hurt performance compared to the original problems. SVAMP (Patel et al., 2021) generated a perturbed benchmark for word math problems, whereas Lila (Mishra et al., 2022) contained perturbations across a diverse range of reasoning questions. The functional variant of the MATH benchmark (Srivastava et al., 2024) demonstrated large performance drops across models when varying existing problems. Putnam-AXIOM (Gulati et al., 2024) similarly shows that perturbing Putnam questions causes a significant drop in model performance. MATH-Perturb (Huang et al., 2025) also adds simple perturbations to math questions (Hendrycks et al., 2021), and shows model performance drops, raising concerns about memorization. Lightman et al. (2024) propose an alternative strategy to improve model robustness by supervising the reasoning process from start to finish, rather than solely on the final outcome. This approach led to improved performance on the MATH dataset. *IMO-Bench* contributes to robust mathematical reasoning with already modified questions in *IMO-AnswerBench*, rigorous proof requirements in *IMO-ProofBench*, and the task of proof grading in *IMO-GradingBench*.

7 Conclusion

This paper introduced *IMO-Bench*, a comprehensive suite of benchmarks for robust evaluation of mathematical reasoning capabilities, including *IMO-AnswerBench* for short answer matching, *IMO-ProofBench* for full proof correctness, and

IMO-GradingBench for proof verification. The results demonstrated that frontier models struggle on *IMO-Bench* problems and that getting the short answers right does not necessarily equate to correct mathematical reasoning for most models.

Furthermore, we have developed and validated automated graders for both answers and proofs. Our *AnswerAutoGrader* achieves near-human accuracy (98.9%) , while *ProofAutoGrader* shows a strong correlation with expert human scores. These tools along with *IMO-GradingBench* provide a scalable and reliable method for the community to evaluate future models, even as human expertise remains the gold standard for high-stakes evaluation.

By releasing *IMO-Bench* to the research community, we aim to shift the community’s focus from mere answer-getting to the development of deep, verifiable, and robust reasoning processes. We hope this suite will serve as a valuable tool to measure and drive progress toward more advanced and reliable artificial intelligence.

Acknowledgments

We thank Miroslav Olšák, Seongbin Jeon, Donghyun Kim, Jiwon Kang, Chu-Lan Kao, Sara Javanmardi, and Mahan Malihi for help with *IMO-Bench*. We thank Orhan Firat, Tania Bedrax-Weiss, and Ed Chi for reviewing the work. We thank Koryay Kavukcuoglu for guidance on the release of *IMO-Bench*. Last but not least, we thank all our collaborators in the IMO 2025 effort⁵ for trusting *IMO-Bench* as north-star metrics along the way.

References

Anthropic. 2025. Introducing claude 4. <https://www.anthropic.com/news/clause-4>.

Evan Chen. 2023. Guidance for problem captains: Or: how to write an olympiad rubric. <https://web.evanchen.cc/static/usemo/captain-guidance-usemo.pdf>.

Yuri Chervonyi, Trieu H. Trinh, Miroslav Olšák, Xiaomeng Yang, Hoang Nguyen, Marcelo Menegali, Junehyuk Jung, Vikas Verma, Quoc V. Le, and Thang Luong. 2025. Gold-medalist performance in solving olympiad geometry with alphageometry2. *Preprint*, arXiv:2502.03544.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reichiro Nakano, Christopher Hesse, and John Schulman.

⁵<https://goo.gle/imo-gold>

2021. Training verifiers to solve math word problems.

Deep Think team. 2025. Try deep think in the gemini app. <https://blog.google/products/gemini/gemini-2-5-deep-think/>.

DeepSeek. 2025a. Deepseek-r1-0528 release. <https://api-docs.deepseek.com/news/news250528>.

DeepSeek. 2025b. Deepseek-v3-0324 release. <https://api-docs.deepseek.com/news/news250325>.

Jingxuan Fan, Sarah Martinson, Erik Y Wang, Kaylie Hausknecht, Jonah Brenner, Danxian Liu, Nianli Peng, Corey Wang, and Michael P Brenner. 2024. Hardmath: A benchmark dataset for challenging problems in applied mathematics. *arXiv preprint arXiv:2410.09988*.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, and 1 others. 2024. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, and 1 others. 2024. FrontierMath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*.

Google DeepMind. 2025. Gemini 2.5 pro. <https://deepmind.google/models/gemini/pro/>.

Aryan Gulati, Brando Miranda, Eric Chen, Emily Xia, Kai Fronsdal, Bruno de Moraes Dumont, and Sanmi Koyejo. 2024. Putnam-axiom: A functional and static benchmark for measuring higher level mathematical reasoning. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Simeng Han, Stephen Xia, Grant Zhang, Howard Dai, Chen Liu, Lichang Chen, Hoang Huy Nguyen, Hongyuan Mei, Jiayuan Mao, and R. Thomas McCoy. 2025. Creativity or brute force? using brainteasers as a window into the problem-solving abilities of large language models. *Preprint, arXiv:2505.10844*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset.

Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jia-wei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, and 1 others. 2025. Mathperturb: Benchmarking llms' math reasoning abilities against hard perturbations. *arXiv preprint arXiv:2502.06453*.

Yichen Huang and Lin F. Yang. 2025. Gemini 2.5 pro capable of winning gold at imo 2025. *Preprint, arXiv:2507.15855*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Thang Luong and Edward Lockhart. 2025. Advanced version of gemini with deep think officially achieves gold-medal standard at the international mathematical olympiad. <https://goo.gle/imo-gold>.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and 1 others. 2022. Lila: A unified benchmark for mathematical reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5807–5832.

Moonshot AI. 2025. Kimi k2: Open agentic intelligence. <https://moonshotai.github.io/Kimi-K2/>.

OpenAI. 2025a. Introducing gpt-5. <https://openai.com/index/introducing-gpt-5/>.

OpenAI. 2025b. Introducing openai o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini/>.

Arkil Patel, Satwik Bhattacharya, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, and 1 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Qwen Team. 2025. Qwen3: Think deeper, act faster. <https://qwenlm.github.io/blog/qwen3/>.

Saurabh Srivastava, Anto PV, Shashank Menon, Ajay Sukumar, Alan Philipose, Stevin Prince, Sooraj Thomas, and 1 others. 2024. Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap. *arXiv preprint arXiv:2402.19450*.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

xAI. 2025. Grok 4. <https://x.ai/news/grok-4>.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. 2021. Minif2f: a cross-system benchmark for formal olympiad-level mathematics. *arXiv preprint arXiv:2109.00110*.

Limitations

Our work has two primary limitations: evaluation cost and the risk of data contamination.

Evaluation Cost. While our automatic grader, *ProofAutoGrader*, correlates strongly with human scores, it is not a perfect substitute and can introduce noise. Consequently, definitive assessments still require verification by human experts, who are both costly and difficult to source.

Future Data Contamination. The second limitation is the risk of long-term data contamination. As *IMO-Bench* is publicly released, its problems and solutions will likely be scraped and absorbed into future training datasets. This threatens the integrity of the benchmark, as models may achieve high scores by memorizing answers rather than demonstrating genuine reasoning. Preventing this form of benchmark decay remains a significant, field-wide challenge.

A IMO-AnswerBench

A.1 Examples

We show examples of *IMO-AnswerBench* in Table 8.

Subj.	Source	Original	Robustified
A	Austria MO 2017	<i>Determine the maximum M of $x+y+z$ where x, y and z are positive real numbers with</i> $16xyz = (x+y)^2(x+z)^2.$	<i>Let a, b, c be lengths of the sides of some triangle of positive area, satisfying</i> $a^2b^2 = 2(a+b-c)(b+c-a)(c+a-b).$ <i>Find the maximum value for $a+b+c$.</i>
C	USA TST 2005	Let n be an integer greater than 1. For a positive integer m , let $S_m = \{1, 2, \dots, mn\}$. Suppose that there exists a $2n$ -element set T such that (a) each element of T is an m -element subset of S_m ; (b) each pair of elements of T shares at most one common element; and (c) each element of S_m is contained in exactly two elements of T . Determine the maximum possible value of m in terms of n .	For a positive integer m , let $S_m = \{1, 2, \dots, 25m\}$. Suppose that there exists a $\underbrace{50\text{-element}}_{\text{Modify numerical value}}$ set T such that: 1. Each element of T is an m -element subset of S_m ; 2. Each pair of elements of T shares at most one common element; 3. Each element of S_m is contained in exactly two elements of T . Let P be a set of 50 random integers. Suppose we define a function $f(x) = \underbrace{x^2 + 2x + 1}_{\text{Add distractors}}$. Determine the maximum possible value of m .
G	USA TST 2024	Let ABC be a triangle with incenter I . Let segment AI intersect the incircle of triangle ABC at point D . Suppose that line BD is perpendicular to line AC . Let P be a point such that $\angle BPA = \angle PAI = 90^\circ$. Point Q lies on segment BD such that the circumcircle of triangle ABQ is tangent to line BI . Point X lies on line PQ such that $\angle IAX = \angle XAC$. Prove that $\angle AXP = 45^\circ$.	Let XYZ be a triangle with incenter J . Let segment XJ meets the incircle of triangle XYZ at point K . Suppose that the angle created by line YK and line XZ is 90° . Let R be a point such that $\angle YRX = \angle RXJ = 90^\circ$. Point S lies on segment YK such that the circumcircle of triangle YSX is tangent to line YJ . Point T lies on line RS such that $\angle JXT = \angle TXZ$. Let γ be the value of $\angle XTR$ in terms of degree, $\underbrace{\text{compute } \frac{\gamma}{3}}_{\text{compute instead prove}}$.
N	Czech- Slovak Math Olympiad 2017	Let $k \neq 0$ be an integer and suppose that there the number of ordered pairs (x, y) of integers satisfying $k = \frac{x^2 - xy + 2y^2}{x + y}$ is odd. Find all possible values of k .	Find all even integers d such that the number of ordered integer pairs (x, y) satisfying $(x + 2y - d)^2 = xy$ $\text{substitute } x \leftarrow x + y, y \leftarrow k - y, d \leftarrow 2k$ is even.

Table 8: Examples in the *IMO-AnswerBench*, per category (Algebra, Combinatorics, Geometry, Number Theory).

A.2 Subject Distribution and Robustification Examples of IMO-AnswerBench

At the IMO, the problems are typically classified into four main categories: Algebra, Combinatorics, Geometry and Number Theory. Therefore, we also structure our *IMO-AnswerBench* in accordance to these four categories as well, where each category has exactly 100 problems.

Algebra is one of the core competencies for Math Olympiad students and appears at all levels of competitions. Distinct from previous benchmarks (Hendrycks et al., 2021), *IMO-Bench* puts more emphasis on Math Olympiad topics, including inequalities, polynomials (including polynomial equations and factorization), functional equations, sequence problems and advanced topics such as Algebraic Number Theory.

Combinatorics problems, despite requiring seemingly basic insights, are notoriously challenging. Successfully solving them serves as a strong indicator of a model’s reasoning capabilities. The combinatorics set of this benchmark contains problems covering Graph Theory, Enumerative Combinatorics (combinatorial counting problems), Extremal Combinatorics, Existence Combinatorics (problems asking the existence of certain combinatorial objects), Additive Combinatorics, Set Combinatorics, Tiling, Combinatorial Geometry, Operations (problems involving operations, often requiring finding invariant or monovariant properties), and Game Theory.

Geometry problems at the IMO are well-known for their visual elegance. While there are several existing geometry benchmarks (Hendrycks et al., 2021), they do not cover Math Olympiad level problems. To address this discrepancy, *IMO-Bench* contains geometry problems with short answers spanning subcategories such as angle and sidelength computation, locus problems, and proof-based geometry problems, as well as unconventional categories such as 3D geometry and combinatorial geometry. Additionally, we would like to note that most Math Olympiad level geometry problems are proof-based, and so designing a Math Olympiad level short-answer benchmark for geometry is highly non-trivial.

Number Theory problems typically consist of problems involving objects and properties derived from integers and arithmetic functions, spanning various topics such as Diophantine equations, divisibility problems, polynomials, sequence problems, functional equation problems on the set of integer, existence problems, problems involving arithmetic functions (such as divisor functions, fractional functions), set problems, number theoretic game problems and strategies such as modular analysis, divisor analysis and base representation problems.

These problems serve as a good representation of Math Olympiad problems at various levels and across different national, regional and international contests, as well as the topics covered in these contests. A strong model performance would suggest a high competence level as well as a good knowledge coverage since certain problems can only be solved with a particular problem solving strategy, without which the model would struggle to provide a rigorous with the correct answer.

A.3 Effects of robustification

To examine the effect of robustification for *IMO-AnswerBench*, we also evaluate on the original, unmodified problems and present the results in Table 9. The models perform significantly better on the original problems, where the gap could be as high as 11.2% for o4-mini (*high reasoning*). This indicates that our robustification effort does create a significant challenge for the models.

A.4 Towards Consistent Problem Statements and Answer Evaluation

Another common issue with language models solving complex Math Olympiad problems is that these models often misinterpret the statement of such problems, or the problem formulation leads the models to produce unintended outputs. Thus, we employ several additional strategies on top of robustification to ensure that the models can interpret the problems properly as follows.

- Instead of asking for a series of numbers satisfying certain conditions (which is hard to verify), we instead reformulate the problem so that its answer is a unique number that is the sum or some other non-trivial function of many inputs.
- Simplifying the answer as much as possible to avoid confusion.

Model	Algebra	Combinatorics	Geometry	Number Theory	Overall accuracy
Grok 4 (Original)	78.8%	61.8%	81.4%	78.3%	75.0% (+1.9)
Grok 4 (Robustified)	75.5%	55.9%	80.1%	80.9%	73.1%
Gemini 2.5 Pro (Original)	77.8%	53.0%	77.4%	78.8%	71.7% (+3.5)
Gemini 2.5 Pro (Robustified)	73.4%	48.0%	74.2%	77.1%	68.2%
o4-mini (high reasoning) (Original)	75.1%	52.9%	82.5%	75.1%	71.4% (+3.5)
o4-mini (high reasoning) (Robustified)	71.2%	46.6%	78.4%	75.3%	67.9%
o3(Original)	67.4%	46.8%	74.1%	67.5%	63.9% (+2.8)
o3 (Robustified)	62.8%	43.0%	70.6%	68.0%	61.1%
Claude Sonnet 4 (Original)	28.2%	15.5%	27.6%	27.6%	24.8% (+1.8)
Claude Sonnet 4 (Robustified)	20.6%	17.8%	26.0%	27.6%	23.0%

Table 9: Comparison between *IMO-AnswerBench* results (Robustified) and results for *IMO-AnswerBench* before robustification (Original). Results are averaged over 8 samples.

- Being more specific with the problem statement to excuse possible issues with special characters, such as angle degrees in geometry problems.
- Avoiding questions with binary answers (yes/no), such as existence questions (which are extremely common in Math Olympiad contests), as they can be guessed without solving the problem or proving the result rigorously. Instead, we will reformulate the problem in such a way that it would produce a non-trivial answer.

A.4.1 Ensuring unique non-trivial answer

Example 1 In this example, instead of asking the model to characterize all such numbers m , we ask the model to compute a certain expression, which results in 1012, a value that the model is unlikely to guess by mere chance.

Original problem: "For a positive integer m , let a_1, a_2, \dots, a_{m+1} satisfy $3^i < a_i < 3^{i+1}$ for each i . Find the maximum and minimum possible values of

$$\sum_{1 \leq x \leq m+1} \prod_{y \neq x} \frac{a_x a_y - 1}{a_x - a_y}. \quad (1)$$

"

Original answer: "maximum of 0 and minimum of 0 if m is odd, and maximum of 1 and minimum of 1 if m is even."

Modified problem: "For a positive integer m , let a_1, a_2, \dots, a_{m+1} satisfy $3^i < a_i < 3^{i+1}$ for each i . Let

$$A_m = \sum_{1 \leq x \leq m+1} \prod_{y \neq x} \frac{a_x a_y - 1}{a_x - a_y}. \quad (2)$$

Find $\sum_{i=1}^{2025} A_m^2$ "

Modified answer: "1012"

Example 2 In this example, instead of asking the model to characterize all solution tuples, which can be hard to evaluate in the natural language form, we ask the models to compute the sum of the elements.

Original problem: "Let $a_1, a_2, \dots, a_{2025}$ be positive integers such that for each positive integer m ,

$$\left(\left(\sum_{j=1}^{2025} j a_j^n \right) - 1 \right)^{\frac{1}{n+1}}$$

is an integer. Find all possible $(a_1, a_2, \dots, a_{2025})$."

Original answer: “ $(a_1, \dots, a_{2025}) = (1, k, \dots, k)$ with $k = 2 + 3 + \dots + 2025 = 2051324$ ”

Modified problem: “Let $a_1, a_2, \dots, a_{2025}$ be positive integers such that for each positive integer m ,

$$\left(\left(\sum_{j=1}^{2025} j a_j^n \right) - 1 \right)^{\frac{1}{n+1}}$$

is an integer. Find all possible values of $a_1 + a_2 + \dots + a_{2025}$.”

Modified answer: “4151879777”

Example 3 In this example, instead of asking the model to characterize all such numbers m , we ask the models to *count* the number of such numbers in a certain range, which results in 1009, a value that the model is unlikely to guess by mere chance.

Original problem: “Find all positive integers $m \geq 2$ that satisfy the following condition: For any m distinct positive integers (n_1, \dots, n_m) , at least one of the following two conditions holds: $n_1 + \dots + n_m$ is a multiple of m , or there exists a permutation (k_1, \dots, k_m) such that $k_1 + 2k_2 + \dots + mk_m$ is a multiple of m .”

Original answer: “All powers of 2 and all odd numbers”

Modified problem: “Find the number of all positive integers $2 \leq m \leq 2000$ that satisfy the following condition: For any m distinct positive integers (n_1, \dots, n_m) , at least one of the following two conditions holds: $n_1 + \dots + n_m$ is a multiple of m , or there exists a permutation (k_1, \dots, k_m) such that $k_1 + 2k_2 + \dots + mk_m$ is a multiple of m .”

Modified answer: “1009”

A.4.2 Answer simplification

Example In the example below the original answer mixes notations and adds a potentially confusing quantifier, so we simplify it.

Original Problem: “Let P be a function from the set of integers to itself such that for all integers h, m , $P^{h^2+m^2}(h+m-1) = mP(m-1) + hP(h-1) + (h+m-1)$. Find all possible functions P .”

Original answer: “ $P \equiv -1$ or $P(x) = x + 1$ for all $x \in \mathbb{Z}$.”

Modified/simplified answer: “ $P(x) = -1, P(x) = x + 1$ ”

A.4.3 Handling geometric quantities

Example For geometry, if the model is asked to find an angle, we make sure to include “in degrees” in the problem statement. This prevents problems mixing radians and degrees and model misinterpretations of special characters marking degrees such as “^o” or “^{\circ}”.

A.4.4 Reformulating questions with binary answers

Example Below is an interesting example where the problem is very difficult but the answer is a binary yes/no, which can be guessed without solving the problem. Therefore, instead of asking the model to show existence, we ask the model to find the smallest positive integer to satisfy such a property, which retains the difficulty level while preventing the model from guessing the answer.

Original problem: “Is there a positive integer n such that $\frac{(a+b)(b+c)(c+a)+a+b+c}{abc} = n$ for infinitely many integer triples (a, b, c) ?”

Original answer: “It exists”

Modified problem: “Find the smallest positive integer n such that there exists infinitely many triple (a, b, c) of distinct positive integers such that $\frac{(a+b)(b+c)(c+a)+a+b+c}{4abc} = n$.”

Modified answer: “3”

A.5 Query prompt for AnswerAutoGrader

The following prompt was used to query the *AnswerAutoGrader* for *IMO-AnswerBench*

```
# System Role: Deterministic Mathematical Autograder
```

You are a precise, automated grading system. Your sole function is to determine if the final answer provided in the Model Solution is mathematically equivalent to the Golden Answer. You must NOT grade the reasoning or steps, only the final result.

1. Grading Guidelines (Equivalence Rules)

Equivalence is mandatory for a correct grade. You must rigorously verify if the answers represent the exact same mathematical value or expression, even if the format differs.

- * **Algebraic Equivalence:** e.g., ' $n(n+1)/2$ ' is equivalent to ' $n^2/2 + n/2$ '. You must verify the algebra.
- * **Numerical Equivalence:** e.g., ' $1/2$ ' is equivalent to ' 0.5 '; ' $\sqrt{2}/2$ ' is equivalent to ' $1/\sqrt{2}$ '.
- * **Set/List Equivalence:** Unless specified as an ordered tuple/vector, the order of elements does not matter (e.g., $\{1, 2\}$ is equivalent to $\{2, 1\}$).
- * **Partial Credit:** No partial credit is allowed. If the answer is incomplete or partially incorrect, it is incorrect.
- * **No Answers:** If no clear, unambiguous final answer can be extracted, the solution must be graded as incorrect.

3. Output Protocol (Strict Compliance Required)

You must execute the task using a two-part structure. Failure to follow this structure will result in task failure.

Part 1: Analysis (Chain-of-Thought)

You MUST perform your analysis within `<thinking></thinking>` tags. Make your thinking concise. This section details your reasoning process and must follow these steps sequentially:

1. **Golden Answer:** State the Golden Answer.
2. **Extracted Model Answer:** State the extracted answer based on the Extraction Protocol. If none found, state "No clear final answer found."
3. **Equivalence Analysis:** Compare the two answers using the Grading Guidelines. Detail the steps taken to verify mathematical equivalence (e.g., simplification, algebraic manipulation). You must actively try to prove they are the same before concluding they are different.
4. **Conclusion:** State the final determination ("Correct" or "Incorrect").

Part 2: Final Grade

Immediately following the closing `</thinking>` tag, output **ONLY** the final grade.

- * If Correct: `\boxed{Correct}`
- * If Incorrect: `\boxed{Incorrect}`

CRITICAL CONSTRAINT: Do not add any text, explanations, or formatting outside the `<thinking>` tags or the final `\boxed{}` output.

Output example:

`<thinking>`

1. **Golden Answer:** $(-\infty, -4) \cup (-4, \infty)$
2. **Extracted Model Answer:** \emptyset (the empty set)

3. ****Equivalence Analysis:****

The Golden Answer is a non-empty set of real numbers. The Model Answer is the empty set. These two sets are not equivalent. The empty set contains no elements, while the Golden Answer contains an infinite number of elements.

4. ****Conclusion:**** Incorrect

```
</thinking>
\boxed{Incorrect}
```

4. Input Data

Here is the problem, model solution, and golden answer to grade:

Problem: {Problem_Statement}
Model Solution: {Model_Solution}
Golden Answer: {Golden_Answer}

B IMO-ProofBench

B.1 Examples

We show robustified examples of *IMO-ProofBench* in Table 10.

B.2 Proof Evaluation Guidelines for IMO-ProofBench

In a proof-based problem, the desired conclusion usually is either already given ("Prove that ...") or easy to guess ("Determine with proof whether ..."). Evaluating a solution consists of verifying that each logical step leading to the conclusion is valid. However, grading informal⁶ proofs contains inherently subjective elements, such as deciding whether a particular claim is justified in sufficient detail. Thus, unlike for short answers, which are either correct or incorrect, it is more appropriate to evaluate proofs on a higher resolution scale, where subjective elements matter less. Additionally, a solution may make partial progress by proving some but not all of the steps of a full solution. It is important to capture this during evaluation.

Traditionally, proof-based Math Olympiad competitions, such as the IMO, score solutions on a 7-point scale. For each problem, a grading rubric outlines how many points are to be awarded for certain partial results. The great majority of solutions receive a polarizing score: either 5-7 points for being essentially correct, or 0-2 points if the problem remains unsolved, generally dictated by specific criteria in the rubric. Although problems often admit multiple solutions, it is rare for a solution to be so novel that it falls completely outside of the rubric (which usually covers the 1-2 most common solution approaches). Thus, despite some elements of subjectivity as mentioned above, scores are typically quite consistent across graders. For further insight into how Math Olympiad grading works, refer to [Chen \(2023\)](#).

B.3 Details of Gemini 2.5 Pro with (Huang & Yang, 2025)

We use the exact agentic framework proposed in [\(Huang and Yang, 2025\)](#), which has been open sourced at <https://github.com/lyang36/IMO25> and also contains exact hyperparameters in its binary flags. We used the same thinking budget (32K tokens) per model call as mentioned in the paper.

Given an initial solution, a single pipeline consists of repeated iterations (at most 30) of "self-verification" and "bug-fixing" on it. Specifically, if the current solution passes self-verification a fixed number (5) of times, then the solution is returned, but if at any time self-verification does not pass, then the model is instructed to observe any mistakes ("bugs") and fix them, and restart the self-verification process.

Note that if a consecutive number (10) of verifications fail, then the pipeline exits without a solution. This entire pipeline will be run in parallel multiple times (100) as well, until there is at least one solution returned from any run. Theoretically the model could fail to find any solution after all parallel runs, which occurred for two *IMO-ProofBench* (Advanced) problems.

⁶i.e. written in natural language, as opposed to a formal language such as LEAN.

Source	Original	Robustified
IMO '24 P1	<p>Determine all real numbers α such that, for every positive integer n, the integer</p> $\lfloor \alpha \rfloor + \lfloor 2\alpha \rfloor + \cdots + \lfloor n\alpha \rfloor$ <p>is a multiple of n. (Note that $\lfloor z \rfloor$ denotes the greatest integer less than or equal to z. For example, $\lfloor -\pi \rfloor = -4$ and $\lfloor 2 \rfloor = \lfloor 2.9 \rfloor = 2$.)</p>	<p>For a real number r, let $A(r)$ denote the fractional part of $2r$ in its decimal representation. For a real number r and a positive integer n, define $B(n, r)$ as</p> $B(n, r) = \sum_{k=1}^n A(kr).$ <p>Find all positive real numbers r such that $n(n+1)r - B(n, r)$ is a multiple of n for all positive integers n.</p>
IMO '24 P2	<p>Determine all pairs (a, b) of positive integers for which there exist positive integers g and N such that</p> $\gcd(a^n + b, b^n + a) = g$ <p>holds for all integers $n \geq N$. (Note that $\gcd(x, y)$ denotes the greatest common divisor of integers x and y.)</p>	<p>For given positive integers x and y, we define a sequence $(a_n)_{n \geq 1}$ where a_n is equal to</p> $\gcd\left(x^n + y, (y - x) \left(\sum_{i=0}^{n-1} y^i x^{n-i-1} - 1\right)\right)$ <p>for all $n \in \mathbb{N}$. Find all pairs (x, y) of positive integers such that the limit of the sequence (a_n) exists.</p>
IMO '24 P5	<p>Turbo the snail plays a game on a board with 2024 rows and 2023 columns. There are hidden monsters in 2022 of the cells. Initially, Turbo does not know where any of the monsters are, but he knows that there is exactly one monster in each row except the first row and the last row, and that each column contains at most one monster.</p> <p>Turbo makes a series of attempts to go from the first row to the last row. On each attempt, he chooses to start on any cell in the first row, then repeatedly moves to an adjacent cell sharing a common side. (He is allowed to return to a previously visited cell.) If he reaches a cell with a monster, his attempt ends and he is transported back to the first row to start a new attempt. The monsters do not move, and Turbo remembers whether or not each cell he has visited contains a monster. If he reaches any cell in the last row, his attempt ends and the game is over.</p> <p>Determine the minimum value of n for which Turbo has a strategy that guarantees reaching the last row on the n-th attempt or earlier, regardless of the locations of the monsters.</p>	<p>On a table of size 3002×3001, a stone is placed on the leftmost cell of the first row. James and Peter play a game on this table. Peter selects 3000 cells, under the rule that he must choose one from each row except the first and last rows (i.e., the 1st and 3002nd row), and there must be at most one selected cell in each column. James knows this rule too, but he doesn't know which cells Peter selected. The goal of James is to move the stone to the last row, avoiding the cells selected by Peter. The stone can only move to adjacent cells on the table. If the stone enters a cell selected by Peter, James receives a penalty of 1 point, and the stone returns to its initial position (i.e., the leftmost cell). Find the smallest positive integer n such that there exists a method for James to achieve his goal before receiving a penalty of n points.</p>

Table 10: Examples of robustified problems, based on the IMO 2024 competition, for *IMO-ProofBench*.

B.4 Common Model Mistakes

Here we list some common mistakes the models were making according to the graders.

B.4.1 Polynomial Assumption

Problem	Model Solution
<p>PB-Basic-004: We want to find all strictly increasing functions $\mathbb{R} \rightarrow \mathbb{R}$ such that:</p> <ol style="list-style-type: none"> 1. g is surjective. 2. $g(g(x)) = g(x) + 20x, \forall x \in \mathbb{R}$. 	<p>First, let us restate the problem clearly: We want to find all strictly increasing functions $g : \mathbb{R} \rightarrow \mathbb{R}$ such that: 1) g is surjective, 2) $g(g(x)) = g(x) + 20x$, for all $x \in \mathbb{R}$.</p> <ol style="list-style-type: none"> 1. Guess and check for linearity <p>A common approach for such functional equations is to hypothesize that g might be linear (i.e., of the form $g(x) = ax + b$). Let us try this:</p> <p>Assume $g(x) = ax + b$, where $a, b \in \mathbb{R}$.</p> <p>Compute $g(g(x))$:</p> $g(g(x)) = g(ax + b) = a(ax + b) + b = a^2x + ab + b.$ <p>According to the functional equation, we need: $g(g(x)) = g(x) + 20x$. Hence, $a^2x + ab + b = (ax + b) + 20x = ax + 20x + b$.</p> <p>Matching coefficients:</p> <ul style="list-style-type: none"> • Coefficient of x: $a^2 = a + 20$. • Constant term: $ab + b = b$. <p>From the constant term $ab + b = b$, we can factor out b: $b(a + 1) = b \Rightarrow$ either $b = 0$ or $a + 1 = 1$.</p> <p>If $a + 1 = 1$, then $a = 0$, but that would force $g(x) = b$, a constant function, which cannot be strictly increasing nor surjective over all \mathbb{R}. Hence, that is not a valid solution.</p> <p>Therefore, we must have $b = 0$. Now the equation $a^2 = a + 20$ remains. Rewrite it as $a^2 - a - 20 = 0$.</p> <p>Factor or use the quadratic formula: $(a - 5)(a + 4) = 0$. This gives $a = 5$ or $a = -4$.</p> <ol style="list-style-type: none"> 2. Strictly increasing and surjective requirement <p>Since g must be strictly increasing:</p> <ul style="list-style-type: none"> • If $a = -4$, then g would be strictly decreasing, which is not allowed. • Hence a must be positive, namely $a = 5$. <p>So the only candidate under our linear assumption is $g(x) = 5x$.</p> <p>Check surjectivity: The map $x \rightarrow 5x$ is surjective, since for any $y \in \mathbb{R}$, we can choose $x = y/5$ to get $5x = y$. It is also strictly increasing because $5 > 0$.</p> <ol style="list-style-type: none"> 3. Verification in the original equation <p>Finally, verify $g(x) = 5x$ satisfies $g(g(x)) = g(x) + 20x$:</p> <ul style="list-style-type: none"> • $g(g(x)) = g(5x) = 5(5x) = 25x$. • $g(x) + 20x = 5x + 20x = 25x$. <p>These match, so $g(x) = 5x$ indeed satisfies both the functional equation and the given conditions.</p> <p>Conclusion</p> <p>Therefore, the only strictly increasing surjective function $g : \mathbb{R} \rightarrow \mathbb{R}$ satisfying $g(g(x)) = g(x) + 20x$ is $g(x) = 5x$.</p>
<p>PB-Basic-005: Let P be a polynomial with real coefficients whose leading coefficient is 1. Suppose that for all nonzero real numbers x, we have $P(x) + P(1/x) = \frac{P(x+1/x)+P(x-1/x)}{2}$. Determine all possibilities for P.</p>	<p>We consider polynomials of degree 2 and 4. For degree 2, we found $P(x) = x^2$ is a solution. For degree 4, we found $P(x) = x^4 + bx^2 + cx + 6$. We found $c = 0$. So $P(x) = x^4 + bx^2 + 6$ is a solution for any real number b. We verified that $P(x) = x^2$ and $P(x) = x^4 + bx^2 + 6$ satisfy the given condition. We showed that polynomials of odd degree are not solutions, and polynomials of degree 0 and 1 are not solutions. We need to show that there are no other solutions. Assume degree of P is n. If n is odd, we showed no solution. So n must be even. Let $n = 2k$. We have shown that for $n = 2$, $P(x) = x^2$ and for $n = 4$, $P(x) = x^4 + bx^2 + 6$. Final Answer: The final answer is $P(x) = x^2$ or $P(x) = x^4 + bx^2 + 6$ for some real number b.</p>

Table 11: Examples of common mistakes that the models made for *IMO-ProofBench*.

PB-Basic-004: We want to find all strictly increasing functions $\mathbb{R} \rightarrow \mathbb{R}$ such that:

1. g is surjective.
2. $g(g(x)) = g(x) + 20x, \forall x \in \mathbb{R}$.

A common mistake that language models typically make while solving problems is assuming strong assumptions on the problem without a legitimate or substantial justification for the assumptions. An instance of this behavior is the functional equation problem PB-Basic-004 as shown in Table 11. The

proof of o3 on this problem begins by considering the linear cases.

"We wish to find all strictly increasing and surjective functions $g : \mathbb{R} \rightarrow \mathbb{R}$ satisfying $g(g(x)) = g(x) + 20x$ for all $x \in \mathbb{R}$.

A natural first step is to check if a linear function works. Suppose

$$g(x) = ax + b$$

...."

After figuring out $a = 5$ and $b = 0$, the model claims that it found the unique solution, even though the only cases it checked were when g is linear.

....

Thus, the unique solution is $g(x) = 5x$."

While the final answer is indeed correct, which a typical short answer benchmark would consider correct, the proof is not rigorous and would get little to no points in a proof-based competition such as the IMO.

B.4.2 Final Answer Guessing

PB-Basic-005: Let P be a polynomial with real coefficients whose leading coefficient is 1. Suppose that for all nonzero real numbers x , we have $P(x) + P(1/x) = \frac{P(x+1/x) + P(x-1/x)}{2}$. Determine all possibilities for P .

In addition, there are the examples where models try to guess the final answer by inspecting the cases when the variables are small. They do not try to actually prove why the guessed answer is correct. In the example problem PB-Basic-005, the model does case work with degree $n = 2$ and degree $n = 4$ and guesses the answer is $P(x) = x^2$ and $P(x) = x^4 + ax^2 + b$ without showing these are correct answers (in fact, the correct answer should have been $P(x) = a(x^4 + 6) + bx^2$) nor that these are all the answers. That being said, the models often can get a lot of correct answers by simply guessing rather than carrying out elaborate derivations to arrive at the correct answer. For more information, we refer the readers to the full example in Table 11.

B.4.3 Commonly Missed Easy Problems

Among many problems that models were not able to solve, we present here the following two pre-IMO difficulty problems from ProofBench-basic.

PB-Basic-008, (Modified) All-Russia MO 2004: Let a, b, c be positive reals such that $a + b + c = 1$, prove that $\sqrt{a} + \sqrt{b} + \sqrt{c} \geq 3\sqrt{3}(ab + bc + ca)$.

This problem is a standard symmetric homogeneous inequality in three variables, whose equality condition is $a = b = c$. This is one of the easiest type of inequalities one could encounter in a high school level math competition. However, not a single model we tested got even a partial score on this one.

PB-Basic-016, (Modified) USAMO 1994 Problem 2: 101 stones are arranged in a circle, numbered 1 through 101 in order (so that stone 1 is next to stone 101). Each stone is painted either red, white, or blue. Initially, stone 101 is the only blue stone. Among the remaining stones, even-numbered stones are red, and odd-numbered stones are white. We perform a sequence of modifications to the coloring, where in each step, we may choose a stone and repaint it a different color, as long as we ensure that no two adjacent stones are ever the same color. Prove that it is not possible to eventually reach a state where again stone 101 is the only blue stone, but among the remaining stones, all even-numbered stones are white and all odd-numbered stones are red.

This is a standard, easy combinatorics problem that uses the invariance principle. One is required to test various quantities related to the problem, and find an invariant quantity to solve the problem. In this case, the difference between the number of blue followed by white and the number of white followed by blue is invariant. However, no model found any invariant that is related to the problem.

B.5 Query prompt for ProofAutoGrader

The following prompt was used to query *ProofAutoGrader* for *IMO-ProofBench*

You are an expert grader for the International Mathematics Olympiad (IMO). Your task is to evaluate a proposed solution strictly and rigorously. Keep in mind the standards at the IMO are extremely high: only arguments that are logically sound, complete, and precise should be rewarded.

General Scoring Rubric

Scores are assigned on a 0-7 scale. The general guidelines are:

- * ****7 Points (Correct):**** The solution is complete, correct, and fully rigorous. If the submission contains incorrect attempts or lines of reasoning but ultimately presents a complete and correct solution, it should still be awarded full points; the presence of earlier, discarded work does not detract from the final correct proof.
- * ****6 Points (Almost Correct):**** The solution is almost correct with a sound core argument, but contains minor errors in calculation or small gaps in logic. Missing proofs for major components, unjustified claims, or sketchy arguments are ****not**** eligible for 6 points.
- * ****1 Point (Partial Progress):**** The solution demonstrates substantial progress explicitly mentioned in the grading guidelines. Initial observations, reformulating the problem without making substantive headway, or proving partial results not mentioned in the grading guidelines are generally ****not**** eligible for this score.
- * ****0 Points (Incorrect):**** The solution doesn't make substantial progress that is a key step in the full solution or is fundamentally flawed. All partial progress without key results or lacking rigor also fall in this category.

Input Data and Interpretation

You are provided with the following:

1. ****Problem Statement:**** The IMO problem.
2. ****Ground Truth Solution:**** A reference solution. Assume this solution is correct. It demonstrates one valid approach.
3. ****Specific Grading Guidelines:**** Criteria for awarding credit for this specific problem. These guidelines take precedence over the General Scoring Rubric, especially for partial credit.
4. ****Proposed Solution:**** The student submission.

Evaluation Process

You must follow this structured process:

1. ****Analyze References:**** Meticulously read and understand the problem and Ground Truth Solution check the Specific Grading Guidelines. Identify the key steps for a complete solution and the criteria for partial credit.
2. ****Step-by-Step Verification:**** Verify the logical validity and rigor of every step. Identify all flaws, gaps, assumptions, and errors. ****Make sure you fully understand every piece of logic behind each step of the proposed solution, you must be careful for solutions that 'pretend' to be correct.****
3. ****Assess Progress:**** Determine the extent of non-trivial progress made.
4. ****Score Determination:**** Compare the findings against the Specific Grading Guidelines and the General Rubric to determine the final score.

Output Requirements

You must provide your final score in the format <points>N out of 7</points>.

Ensure the ‘`<points>`’ block is used ****only once****, as your answer will be parsed based on the first `<points> </points>` block that appears in your whole response.

****PROBLEM STATEMENT****

{problem_statement}

****GROUND-TRUTH SOLUTION****

{solution}

****SPECIFIC GRADING GUIDELINES****

{guidelines}

****PROPOSED SOLUTION****

{student_answer}

Present your detailed thought process and formal justification based on the scoring rubric and grading guidelines, and finally present your final score in the format below.

[Select one of the following options]

`<points>7 out of 7</points>`
`<points>6 out of 7</points>`
`<points>1 out of 7</points>`
`<points>0 out of 7</points>`

B.6 Limitations of ProofAutoGrader

Despite a high correlation with human grades, *ProofAutoGrader* still has several systematic errors and limitations, including a general tendency to overestimate scores, occasional failure to identify high-level logical errors, and being prone to be overly punitive for minor formatting issues or unconventional yet correct solutions. We demonstrate specific examples of these behavior in Table 12.

In PB-Basic 002, the model solution makes a logical error by asserting $2(4\sqrt[4]{xyzt}) \geq xyzt$ directly from $x + y + z + t \geq 4\sqrt[4]{xyzt}$ and $2(x + y + z + t) \geq xyzt$. This comes from an incorrect assumption that if $A \geq B$ and $A \geq C$, then $B \geq C$. Such “specious” errors, while seemingly plausible and easy to overlook without a deep understanding of the problem, are critical and can invalidate an entire solution. *ProofAutoGrader* often fails to identify such deceptive logical inconsistencies.

In PB-Basic 027, the model produces a novel solution entirely different from the established ground truth and grading guidelines. The solution was largely correct, but its ‘Key Lemma’ omits a critical condition that the segment PQ must have a fixed slope. While the lemma is false as stated, supplying this condition makes its proof an immediate consequence of homothety. Since the rest of the solution is complete, the human grader awarded it 6 out of 7 points. However, because the lemma is technically incorrect, *ProofAutoGrader* marks the entire solution as wrong. This case demonstrates that *ProofAutoGrader* struggles to identify partial progress in solutions not anticipated by the grading guidelines, leading to overly punitive assessments for minor issues.

C IMO-GradingBench

C.1 Grade distribution for IMO-GradingBench

This section presents the human-assigned grade distribution for the *IMO-GradingBench* benchmark. As shown in Figure 6, the aggregate count of correct versus incorrect grades across the entire dataset is balanced.

However, the distribution of grades (correct, almost, partial, incorrect) is not uniform on a per-problem basis. This variance is expected as it reflects the natural distribution of scores that proof-evaluation models will encounter in grading solutions, as problems inherently differ in difficulty.

Problem	Model Solution and Grade
PB-Basic 002: Show that $x^2 + y^2 + z^2 + t^2 \geq xyzt$ for any positive real numbers x, y, z, t that satisfy $2(x + y + z + t) \geq xyzt$.	<p>Model Solution: Of course. This is a great problem that elegantly combines a given condition with a target inequality. The key is to use the given condition to constrain the value of the product $xyzt$. Here is the step-by-step proof.</p> <p>**Step 1: Use the given condition and the AM-GM inequality to find a bound for the product $xyzt$.**</p> <p>We are given two pieces of information for positive real numbers x, y, z, t: 1. The condition: $2(x + y + z + t) \geq xyzt$ 2. The inequality to prove: $x^2 + y^2 + z^2 + t^2 \geq xyzt$</p> <p>Let's start with the Arithmetic Mean-Geometric Mean (AM-GM) inequality for the four numbers x, y, z, t:</p> $\frac{x + y + z + t}{4} \geq \sqrt[4]{xyzt}$ <p>Multiplying both sides by 4, we get:</p> $x + y + z + t \geq 4\sqrt[4]{xyzt}$ <p>Now, we can substitute this result into the given condition $2(x + y + z + t) \geq xyzt$:</p> $2(4\sqrt[4]{xyzt}) \geq xyzt$ $8\sqrt[4]{xyzt} \geq xyzt$ <p>(...)</p> <p>Human Grade: 0 out of 7</p> <p>Autograder Verdict: ### Step-by-Step Verification</p> <p>1. **Step 1: Bounding the product 'xyzt'** * The student starts with the given condition: $2(x + y + z + t) \geq xyzt$. * They correctly apply the AM-GM inequality to x, y, z, t: $x + y + z + t \geq 4\sqrt[4]{xyzt}$. * Substituting this into the given condition yields $2(4\sqrt[4]{xyzt}) \geq xyzt$, which simplifies to $8\sqrt[4]{xyzt} \geq xyzt$.</p> <p>(...)</p> <p>The solution is complete, correct, and fully rigorous. It meets all the criteria for a perfect score according to the General Scoring Rubric. It does not contain any errors, minor or otherwise. The method is different from the provided ground truth solution (it's a direct proof rather than a proof by contradiction), but it is equally valid and elegant.</p> <p><points>7 out of 7</points></p>
PB-Basic 027: Given an acute triangle ABC , let D and E be the feet of the altitudes from B to AC and C to AB , respectively. Let E_1 and E_2 be the reflections of E with respect to AC and BC , respectively. If X (not equal to C) is an intersection point of the circumcircle of $\triangle CE_1E_2$ and AC , and O is the circumcenter of $\triangle CE_1E_2$, prove that XO is perpendicular to DE .	<p>Model Solution: Let H be the orthocenter of triangle ABC. Let H_a and H_b be the reflections of H about AC and BC, respectively; then H_a, H_b lie on the circumcircle (ABC). Note that reflecting the line CH about AC gives the line CH_a, hence E_1 lies on CH_a; similarly E_2 lies on CH_b.</p> <p>Key lemma (fixed tangent direction): Let $l = AC$, $r_1 = CH_a$, $r_2 = CH_b$. For any circle ω through C intersecting l, r_1, r_2 again at X, P, Q respectively, the angle between the tangent to ω at X and XC is constant (independent of the chosen $P \in r_1$ and $Q \in r_2$). Indeed, in the cyclic quadrilateral $XPCQ$ we have:</p> <ul style="list-style-type: none"> • $\angle XCP = \angle(l, r_1)$ and $\angle XCQ = \angle(l, r_2)$, both constants; • $\angle XPC = \angle XQC$ (equal angles subtending chord XC); • hence $2\angle XPC + \angle XCP + \angle XCQ = 180^\circ$, so $\angle XPC = 90^\circ - (\angle(l, r_1) + \angle(l, r_2))/2$ is constant; • by the tangent–chord theorem, this $\angle XPC$ equals the angle between the tangent at X and XC. <p>(...)</p> <p>Human Grade: 6 out of 7</p> <p>Autograder Verdict: (...)</p> <p>### Assessment and Score Determination The solution's overall structure is elegant, but it hinges entirely on a "Key Lemma" which is false and whose proof is invalid. The argument is fundamentally flawed.</p> <p><points>0 out of 7</points></p>

Table 12: Examples of failure cases of *ProofAutoGrader*.

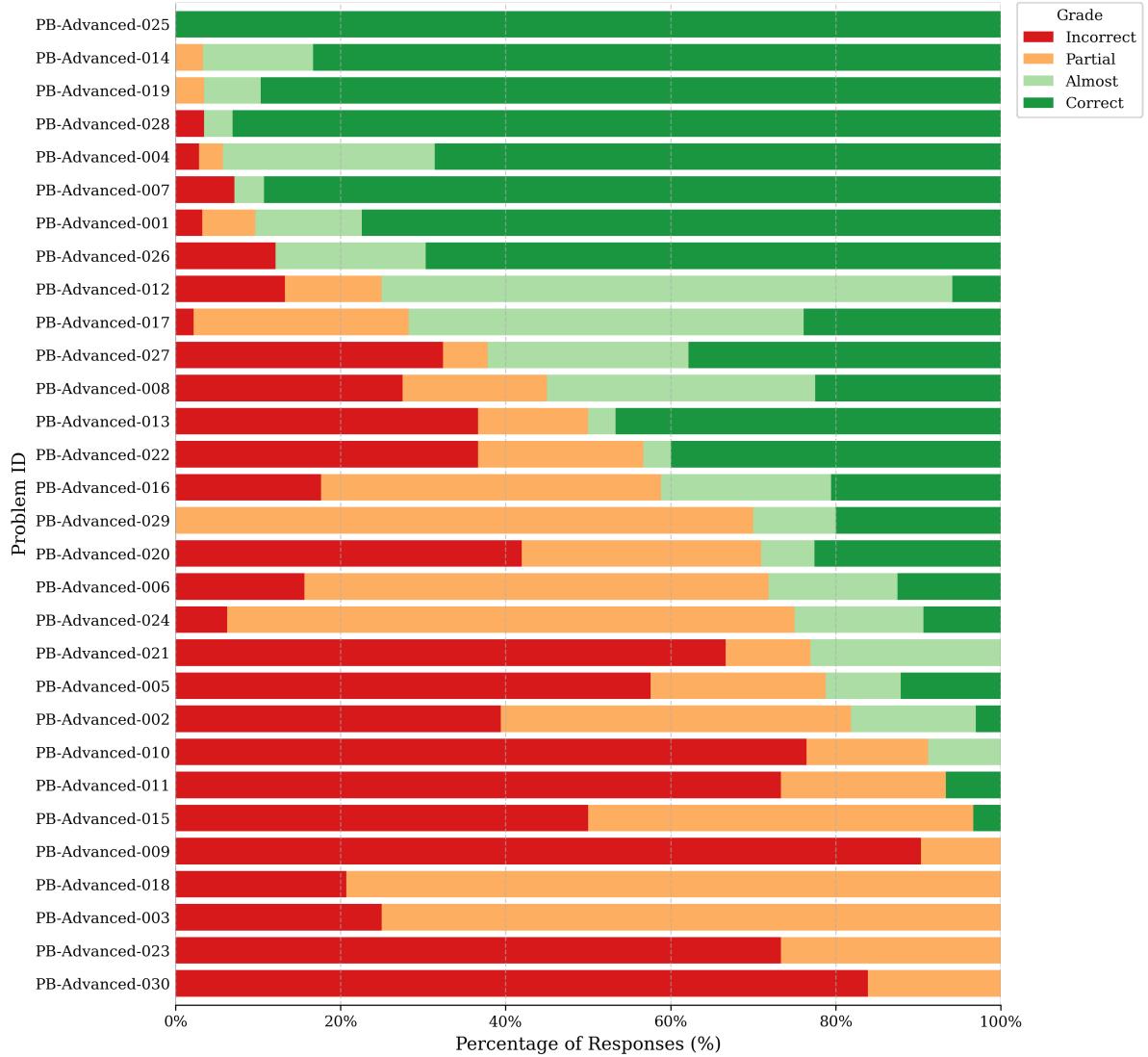


Figure 6: Grade distribution across examples in *IMO-GradingBench*

C.2 Query Prompt

This section details the prompts used for the three evaluation settings in *IMO-GradingBench*. A common definition of the scoring criteria is used across all settings, inserted into the prompts as indicated by `{SCORING_CRITERIA}`.

C.3 Grader Prompt

The following prompt was used for the vanilla setting:

Carefully analyze the given problem statement and the proposed solution, and then write out your analysis regarding the correctness of the proposed solution.

After the analysis, you must provide a score based on the following criteria:

- **incorrect:** The solution is completely incorrect or irrelevant.
- **partial:** The solution is partially correct but has significant errors or omissions.
- **almost:** The solution is almost correct but contains minor errors or inaccuracies.
- **correct:** The solution is fully correct and complete.

The very last part of your response must be **only** one of the following words: incorrect, partial, almost, or correct.

Problem:{problem} Solution:{solution}

C.4 Label extraction prompt

The following prompt was used to extract the label from model response for *IMO-GradingBench*. Note that in the majority of cases, the last word of the model (grader) response is one of incorrect, partial, almost, or correct. As a result, we first use python to extract the model grades. We only use prompting to extract the model grades when the last word in the model response is empty or is some different words.

Instructions for Extracting Final Scores

****Objective:**** Given an response of an evaluation prompt, extract the final score presented within the response and format it specifically.

****Process:****

1. ****Analyze the response:**** Scan the response to identify the final score provided by the evaluator.
2. ****Extract and format the final answer:**** Present the extracted score on a new line, preceded exactly by "Final answer: ".

****Formatting Rules:****

- * ****Evaluation Categories:**** The expected output must be one of the following categories: 'correct', 'partial', 'almost', 'incorrect', or 'not found'.
- * ****Score Identification:**** The extraction is based on identifying the keyword used by the evaluator to summarize their conclusion. The criteria associated with these keywords are:
 - * ****incorrect:**** The evaluator concluded that the solution is completely incorrect or irrelevant.
 - * ****partial:**** The evaluator concluded that the solution is partially correct but has significant errors or omissions.
 - * ****almost:**** The evaluator concluded that the solution is almost correct but contains minor errors or inaccuracies.
 - * ****correct:**** The evaluator concluded that the solution is fully correct and complete.
 - * ****not_found:**** The evaluation response does not clearly contain one of the four explicit scores listed above.
- * ****Extraction:**** Determine the provided score from the response and extract the category ('correct', 'partial', 'almost', or 'incorrect'). If a score cannot be reliably identified within the text, the output must be 'not_found'.

****Note:**** No additional markings or explanations are needed beyond "Final answer: " and the extracted answer.

Below is the response:

{Model Response}